Hindawi
Education Research International
Volume 2021, Article ID 5543317, 15 pages
https://doi.org/10.1155/2021/5543317

Research Article

Psychometric Properties of Student Evaluation of Teachers' Performance Scale: Evidence from Debre Markos University Students' Evaluation Dataset

Muluye Getie Ayaneh,[1] **Askalemariam Adamu Dessie**,[2] **and Dimetros Molla Fetene**[3]

[1]*Department of Statistics, College of Natural and Computational Science, Debre Markos University, P.O. Box 269, Debre Markos, Ethiopia*
[2]*Department of Psychology, Institute of Education and Behavioral Sciences, Debre Markos University, Debre Markos, Ethiopia*
[3]*Debre Markos Teachers Training College, Debre Markos, Ethiopia*

Correspondence should be addressed to Muluye Getie Ayaneh; muluyegb@gmail.com

Received 13 February 2021; Revised 16 June 2021; Accepted 24 June 2021; Published 5 July 2021

Academic Editor: Elina Kuusisto

Introduction. Student evaluation of teachers' effectiveness is one of the most common tools used as a measure of teaching performance and accountability by various universities across the globe. The major purpose of this study was to evaluate the validity and underlying structure of students' evaluation of the higher education teaching effectiveness scale used by all public universities in Ethiopia. *Methodology*. Data collected from 1397 students at Debe Markos University were used for this analysis. Cronbach's alpha values and average interitem correlation were used to study the internal consistency reliability of the scale. Composite reliability, average variance extracted, hetero trait-mono-trait ratio, maximum shared variance, average shared variance, and interconstruct correlations were used to assess the construct validity of the scale, and exploratory factor analysis and confirmatory factor analysis were performed with 20 items to test the hypothesis which introduced a four-dimensional construct for teachers' evaluation scale. We used different goodness-of-fit indices to measure the fit of the models. *Results*. The scale was shown to have good internal consistency and convergent validity but lacked discriminant validity. Furthermore, confirmatory factor analysis indicated that the four-factor model produced inadequate fit indices, revealing that the original factor structure of the scale changed. *Conclusions*. The results showed that Student Evaluation of Teaching Effectiveness did not measure what it was supposed to be measuring. Moreover, the exploratory factor analysis and confirmatory factor analysis results indicate that a two-dimensional model is better than the four-dimensional model to explain the data structure, which places limitations on its use.

1. Introduction

Reliability and validity, jointly called the "psychometric properties" of measurement scales, are the two most important and fundamental features in the evaluation of any measurement scale [1–4]. The evidence of validity and reliability are prerequisites to assure the integrity and quality of a measurement scale [5].

Student evaluation of teachers' effectiveness (SETE) is commonly used to measure teaching performance and accountability by various universities across the globe [6,7]. If carefully developed and systematically used, teacher evaluation is believed to have the potential to enhance teachers' professional development, thereby improving students' achievement [8]. Hence, the scales used to assess teacher performance should be accurate and exhaustive, allowing the results to provide useful information about teachers' teaching effectiveness.

The effectiveness of an education system largely depends on the effectiveness of its teachers, which in turn has a large influence on student learning [9]. As a result, measuring teachers' effectiveness is an important vehicle for promoting educational quality [9–11], which in turn enhances the quality of graduates [12].

Currently, student evaluation of teaching effectiveness is a common practice in almost every institution of higher education globally [13]. Over the years, however, different SETE scales have been proposed and developed. Consequently, numerous well-designed and validated instruments are available to measure higher education teachers' teaching effectiveness [7]such as the Students' Evaluation of Teaching Effectiveness Rating Scale [14], the Student Course Experience Questionnaire [15], the questionnaire for student evaluation of teaching [16], and the Teaching Proficiency Item Pool [17]. The development of the SETE scale is an ongoing process to develop a psychometrically sound scale that measures teacher effectiveness in higher education taking into account the dynamics of the characteristics of effective teaching.

A number of studies have been conducted to address the various elements of SETE scale [7,18,19]. Researchers claim that the SETE scale should capture multiple aspects (dimensions) of good teaching practices [7]. Some of the studies have asserted that the SETE scales need to be one-dimensional [18,20], whereas others believe it to be multi-dimensional [7, 19–31].

The variations in the content and the number of dimensions are attributed to the absence of agreement concerning the number and nature of these dimensions, which should be based on both theory and empirical testing [7]. Identifying the characteristics of effective teaching, which is a prerequisite for the construction of SETE scales, is also a possible reason for the variation in the SETE scale. Moreover, different institutions have different educational visions and policies, thereby developing SETE scales that are consistent with their preferences.

In the search for educational quality in Ethiopia, various attempts have been undertaken to generate meaningful and accurate indices of teacher effectiveness [32]. To this end, the Ministry of Science and Higher Education (formerly Ministry of Education) in Ethiopia identified four competencies of teacher effectiveness that served as the conceptual basis for this study: subject matter knowledge (core competency), professional competency, ethical competency, and time management [33]. The first two indices are related to the instructional effectiveness of teachers, whereas the next two indices are related to the teacher's personal quality. Each of these dimensions focuses on a key aspect of a teacher's professional qualification or responsibilities. As a result, it is critical to determine if the scale measures the intended competences or construct accurately and consistently [1–4].

Nevertheless, no study has been conducted on the psychometric properties and validity of the SETE scale used by Ethiopian higher education institutions despite many faculty members questioning the validity and reliability of SETE results for many years. The scale was not evaluated by independent experts and the target population (students) to verify that the items adequately measure the domain of interest. Pretesting has not been made to assess the extent to which items reflect the constructs of interest. In addition, the SETE scale was not evaluated to test the dimensionality, reliability, and validity. Rather, to the researchers' best knowledge, the factor structure of the scale was constructed merely via discourse between subject matter experts. With the researchers' sufficient experience in the study area, no previous studies have been conducted to investigate the factor structure (dimensionality), internal consistency, and validity of the SETE scale. Therefore, the use of student evaluation of teachers' effectiveness scale claimed to have many problems concerning reliability, validity, dimensionality, and potential bias. Thus, this study was carried out with the major purpose of evaluating the reliability, validity, and underlying factor structure of the SETE scale. More specifically, the study aimed to determine whether the scale could indeed measure the unobservable construct/domain that was supposed to measure or check if the scale revealed an equivalent factor structure with what was established by the experts, and to test the convergent and discriminant validity of students' evaluation of SETE.

Specifically, this search sought to answer the following questions:

Does a SETE scale demonstrate adequate reliability at scale and item levels?

To what extent does the SETE scale show construct and discriminant validity?

Is the factor dimensionality of the SETE scale appropriate to measure higher education teaching effectiveness?

2. Methodology

2.1. Population and Context of the Research. Debre Markos University in Ethiopia is one of the public Universities founded by the Ethiopian Federal government in 2007. The university is located in East Gojjam, Amhara National Regional State, 300 km in northwest of the capital Addis Ababa. Currently, the university runs 51 bachelor's, 47 master's, and 2 Ph.D programs in regular, continuing, and distance education streams. There are more than 1556 academic staff and 1600 administrative staff in the university to serve over 30000 students and the community at large.

In Ethiopian higher education institutions, the application of the SETE is carried out at the end of the semester, before the final exams are administered, and the students know their final grades. All teachers are evaluated by the students in the same semester.

2.2. Local Context of SETE Development Process. The Students' Evaluation of Teaching Effectiveness (SETE) scale Table 1 is one of the three harmonized scales used to measure teachers' effectiveness. These scales were developed by the Ethiopian Ministry of Science and Higher Education. A group of subject matter experts developed the SETE scale, which comprised 20 items and judged the dimensions to be four: subject matter knowledge, professional skills, ethical quality, and time management [34]. From the four constructs, knowledge of the subject matter was considered as the core competence.

Three bodies are involved in assessing teachers' competencies: students, peers, and immediate supervisors [34]. The students' evaluation of teaching effectiveness accounted for 50% of the total evaluation, and the remaining 30% and 20% of the evaluations were accounted for by immediate supervisors and colleagues, respectively. The SETE scale items have five-point Likert scales, of which only one alternative may be chosen. Scores range from 1 to 5, where: 1 = "strongly disagree," 2 = "disagree," 3 = "neutral," 4 = "agree," and 5 = "strongly agree." The 20 items that make up the SETE scale were broadly structured to reflect two teaching effectiveness factors or constructs: 14 for core and professional competency constructs and the remaining four items were related to the ethical and time management constructs.

To ensure the relevance of the items to the general principles of teaching in higher education settings, the development of the scale went through several steps to receive feedback from different stakeholders. To do so, different focus group discussions were held with department heads, students, and college deans.

2.3. The Data and Study Participants. This study is a secondary analysis carried out on data from the teachers' assessment survey, which was undertaken at Debre Markos at the end of every semester to monitor the performance of teachers concerning teaching and research work activities. The following steps were followed to collect (extract) data for this study. In the first step, the teachers to be included in the sample were randomly selected. Then an excel data abstraction tool was prepared to record and manage the teacher's assessment score. To assist with data abstraction and data entry process, a total of 10 data collectors, one from each sampled department were selected. The data collection process was supervised by the quality assurance office of the university. The data were collected anonymously. The evaluation records of 1397students were randomly selected from a population of 5257 regular students who were active in the 2018/2019 academic year. For lower costs and smaller prediction errors, a multistage stratified random sampling was employed to select teachers' evaluation records. We followed the following steps; in the first step, we divided the population of teachers into homogeneous, mutually exclusive subgroups called colleges/faculty. In the second stage, a sample of departments was randomly taken. In the third step, teachers were stratified by their sex. Finally, a sample of teachers was randomly selected for each sex category and then their evaluation records were extracted.

The probability proportional to the size sampling method was used for selecting teachers from each department. Accordingly, 92 (78.6%) of the teachers were male and the remaining 25(21.4%) were females.

2.4. Data Analysis

2.4.1. Reliability and Validity of the Scale. Descriptive measures such as Cronbach's alpha and average interitem correlations were used to assess internal consistency reliability.

We used the CFA method to test the convergent validity, discriminant validity, and nomological validity of a measurement model [35]. Convergent validity measures the extent to which different measures of the same construct converge or strongly correlate with one another, whereas discriminant validity is the extent to which measures of different constructs diverge or minimally correlate with one another [36]. Convergent validity comprises composite reliability (CR) and average variance extracted (AVE). CR, which indicates the shared variance among the observed variables of a latent construct, was applied to test the degree to which the indicator variables converged and shared the proportion of variance [35]. This is calculated using.

$$\mathrm{CR} = \frac{\sum_{i=1}^{p} \lambda_i}{\left(\sum_{i=1}^{p} \lambda_i\right)^2 + \sum_{i=1}^{p} V(\sigma)}, \tag{1}$$

where λ_i is the completely standardized loading for the i^{th} indicator, δ_i is the variance of the error term for the i^{th} indicator, and p is the number of indicators.

Moreover, the average variance extracted (AVE) represents the average amount of variance of constructs, which is explained by its indicator variables relative to the overall variance of its indicators. This is similar to the explained variance in EFA, as it measures the average variance in the items that a construct manages to explain [37]. A higher AVE value indicates lower error variance. The AVE for the j^{th} construct, denoted by C_j is defined using:

$$\mathrm{AVE}_{C_j} = \frac{\sum_{k=1}^{k_j} \lambda_{jk}}{\sum_{k=1}^{k_j} \lambda_{jk}^2 + \theta_{jk}}, \tag{2}$$

where λ_{jk} is the indicator loading and θ_{jk} is the error variance of the k^{th} indicator ($k = 1,\ldots, K_\mathrm{j}$) of the j^{th} construct score (C_j). K_j is the number of indicators of the j^{th} construct C_j. If all indicators are standardized (i.e., having a mean of 0 and a variance of 1), equation (2) simplifies to (3).

$$\mathrm{AVE}_{C_j} = \frac{1}{k} \sum_{k=1}^{k_j} \lambda_{jk}^2. \tag{3}$$

In this case, the AVE is the same as the average squared standardized loading and is equivalent to the mean value of the indicator reliabilities. Now, let r_{ij} be the correlation coefficient between the construct scores of constructs C_i and $C_{j.}$ The squared interconstruct correlation r_{ij}^2 indicates the proportion of variance that constructs C_i and C_j have.

According to the Fornell–Larcker criterion [38], discriminant validity is established if the condition in equation (4) holds.

$$\mathrm{AVE}_{C_j} = \max r_{ij}^2 \text{ for all } i \neq j \Leftrightarrow \sqrt{\mathrm{AVE}_{C_j}} = \max\left|r_{ij}\right| \text{for all } i \neq j. \tag{4}$$

That is, the square root of AVE should be greater than the interconstruct correlations for all constructs. Discriminant validity can also be evaluated using the maximum shared variance (MSV) and average shared variance (ASV), which measure the maximum variance and average variance among constructs, respectively. Both measures should be lower than the AVE for all constructs to confirm discriminant validity [39].

2.4.2. The Heterotrait-Monotrait Ratio Approach. Henseler et al. [39] suggested using the heterotrait-monotrait ratio (HTMT) of correlations, which is the average of the heterotrait-monotrait method correlations (i.e., the correlations of indicators across constructs measuring different phenomena), relative to the average of the mono-trait-hetero method correlations (i.e., the correlations of indicators within the same construct). Because there are two mono-trait-hetero method submatrices, we take the geometric mean of their average correlations. Consequently, the HTMT of constructs Ci and C_j with K_i and K_j indicators can be formulated as .

$$\mathrm{HTMT}_{ij} = \frac{1}{k_i k_j} \sum_{g=1}^{k_i} \sum_{h=1}^{k_i} r_{ig,jh} \div \left(\frac{1}{k_i\left(k_i - 1\right)} \sum_{g=1}^{k_{i-1}} \sum_{h=g+1}^{k_i} r_{ig,ih} \frac{1}{k_j\left(k_j - 1\right)} \sum_{g=1}^{k_{j-1}} \sum_{h=g+1}^{k_j} r_{jg,jh} \right)^{1/2}, \tag{5}$$

where the numerator and the denominator in equation (5) represent the average hetero trait-hetero method and the geometric mean of the average mono-trait-hetero method, the correlation of construct C_i, and the average mono-trait-hetero method correlation of construct C_j, respectively.

2.4.3. Exploratory Factor Analysis. Exploratory factor analysis (EFA) is appropriate when the goal of research is to create a measurement scale that reflects a meaningful underlying construct(s) represented in the observed variables [40]. It is a popular approach to test whether item-level discriminant validity is established by assessing cross-loading [39].

In EFA, the challenge is determining the required number of factors to retain a sufficient amount of variance and, at the same time, to achieve a substantial reduction in dimensionality [41,42]. Several methods are available for determining the number of components or factors for EFA, but they do not always lead to the same or even similar results. Despite the importance of factor retention decisions and extensive research on methods for making retention decisions, there is no consensus on the appropriate criteria to use [43].

2.4.4. Confirmatory Factor Analysis. A confirmatory factor analysis (CFA), which has wide applications in the area of scale development and construct validation [35], was used to determine the validity of the factor structure of the teaching effectiveness assessment scale used by students. Confirmatory factor analysis (CFA) is a popular structural equation model that provides the simplest explanation of how observed and latent variables are related to assumed latent variables [44]. CFA provides a more explicit framework for confirming prior notions about the factor structure of scales [45]. It has two components. The first is a measurement model that explores the relations between a set of observed variables, also called manifest variables (items in our case), to a usually smaller set of latent variables (factors or constructs). The second is a structural model that explores the relationship between latent variables through a series of recursive and nonrecursive relationships. In this study, a four-factor measurement model was specified to test the validity and reliability of the observed indicator items measured on the knowledge of the subject matter (core competency), professional skills (competency), ethical quality, and time management constructs. Professional competency here refers to the degree to which teachers are utilizing their knowledge, skills, and good judgment related to their teaching activities to render tasks with acceptable quality.

Confirmatory factor analysis was carried out using the lavaan package version 0.6–7 [46] in *R* statistical software version for Windows [47]. By examining three critical sets of results—parameter estimates, fit index, and potential modification indices—researchers formally tested the measurement hypotheses, and they can modify the hypotheses to be more consistent with the actual structure of participants' responses to the scale.

3. Results

3.1. Preliminary Data Analysis. Prior to the analysis, we examined missing values and outliers. The missing values of the corresponding variables were imputed by median values. Figure 1 shows a graphical visualization of missing values, which is produced by Visdat package [48]. The figure provides the pattern and percentages of missing value distribution. It also shows the locations of missingness that occurred in the data. From Figure 1, there were 3.2% missing values and 96.8% present values in the dataset. Missing data on item level was low, except items Core5 (6.9%), Core (9.3%), and Ethic15 (7.3%). From the figure, it is apparent that the pattern of missingness is random.

For variables measured on an ordinal scale, neither the assumption of normality nor the continuity property is met [49]. The results presented in Table 2 show that the skewness measures are significantly negative in all items, indicating that maximum values are more common than smaller values. Kurtosis exceeds the reference value of the normal distribution (equal to 3) for the majority of competency components, suggesting the existence of heavy tails compared to the Gaussian distribution. This leptokurtic behavior confirms a typical distribution that exhibited fatter tails than the normal distribution. When the assumption of normality is severely violated, the diagonally weighted least squares (DWLS) method, which is a robust WLS method [50], was used as it provides more accurate parameter estimates [49–51].

3.2. Reliability of the SETE Scale. Table 2 presents the means, standard deviations of items, and internal reliability coefficients for the factors/constructs. Accordingly, all reliability coefficient estimates of alpha except time management skill are above the traditional cutoff of 0.70, revealing that the three teaching competency dimensions/factors have sufficient internal consistency. That is, the reliability of subject matter knowledge (core competency) (Cronbach's alpha = 0.88), professional competency (Cronbach's alpha = 0.89), and ethical competency (Cronbach's alpha = 0.83).

Furthermore, the corrected item-total correlation ranged from 0.44 to 0.85, which exceeded the accepted cutoff of 0.40 proposed by Nunnally [52], indicating that each item was related to the corresponding components of the SETE scale.

In addition, the values of the "reliability if an item is dropped" show a lower or equal value to the alpha value for all variables of the three factors, indicating that all items in all factors contribute positively to the internal consistency of the factors [53]. In Table 2, it is also revealed that the means of the items scale ranged from 3.707 to 4.719, while the standard deviations of the items were from 0.70 to 1.46, indicating a narrow spread around the mean.

The average interitem correlation (AIIC) was computed from the interitem correlation matrix. Correlation matrix presented in Figure 2. The ideal range for the AIIC value is between the values 0.20 and 0.40 [54]. Piedmont (2014) claimed that an AIIC score of less than 0.20 indicates that the items are not well correlated and do not measure the same construct or factor, whereas an AIIC score greater than 0.40 suggests that the items in the same construct are redundant.

Accordingly, the average interitem correlations (AIIC) for core competency, professional competency, ethical quality, and time management skill constructs were 0.58, 0.52, 0.59, and 0.48, respectively, suggesting that all constructs of the SETE scale contain items that measure the constructs in the same way. However, it seems that some of the items in each competency component are redundant.

3.3. Validity of the SETE Scale. The discriminant and convergent validity of the scale were tested using various techniques. The interitem correlation matrix presented in Figure 1 was used for the first visual diagnosis of the items and scale structure. The results displayed in the figure provide evidence that our items in each factor or construct had a high correlation, implying that the items in each construct were related, indicating that the convergent validity of the scale was assured. Composite reliability (CR) and average variance extracted (AVE) were used to test the extent to which the indicator variables converged and shared the proportion of variance. According to Adedeji et al. (2017), a cutoff point of 0.7 or above for CR is required to establish that the indicator items are reliable, and a minimum value of 0.5, which is required for AVE. Furthermore, CR values higher than the AVE are required to establish convergent validity. Accordingly, Tables 3 and 4 present the convergent validity and discriminant validity assessment results. CR values for CC (0.88), PC (0.88), EC(0.85), and TM (0.67) are all above 0.7 (the cutoff point), fulfilling the required threshold. This confirms that convergent validity is established. Moreover, convergent validity is established when CR is higher than AVE, and the AVE is higher than 0.5 [45,55]. These conditions were confirmed in this study; consequently, the convergent validity of the scale was verified. Furthermore, from the lavaan output presented in Table 5, all items appeared to be significantly associated with their respective constructs, which provides additional evidence of convergent validity. Discriminant validity analyzes how well the constructs are distinct and uncorrelated. The scale faces a discriminant validity problem if the items correlate more highly with variables outside their parent factor than with the variables within their parent factor; that is, the latent factor is better explained by some other variables (from a different factor) than by its observed variables. We used the Fornell–Lacker criterion [38], which compares the square root of the average variance extracted (AVE) with the correlation of latent constructs to assess discriminant validity. The interconstruct correlations among the four constructs are shown in Table 3. A strong correlation between them is evidence of their dependence on one another. Accordingly, based on the estimates presented in Table 3, the square root of AVE is less than the interconstruct correlation. Furthermore, the results presented in Table 4 indicate that the maximum shared variance is greater than the average shared variance, and the average shared variance is greater than the average variance extracted (i.e., MSV > AVE and ASV > AVE). Consequently, both results justify the establishment of discriminant validity. The highlighted cells in Table 4 show the HTMT ratio of the correlation between the two constructs, which is calculated using equation (5), as proposed by Henseler et al. [39]. Accordingly, the HTMT values are above the suggested threshold of 0.85 [56], revealing that discriminant validity does not exist between the two reflective constructs, which supports the above finding. In conclusion, the scale faced a discriminant validity problem.

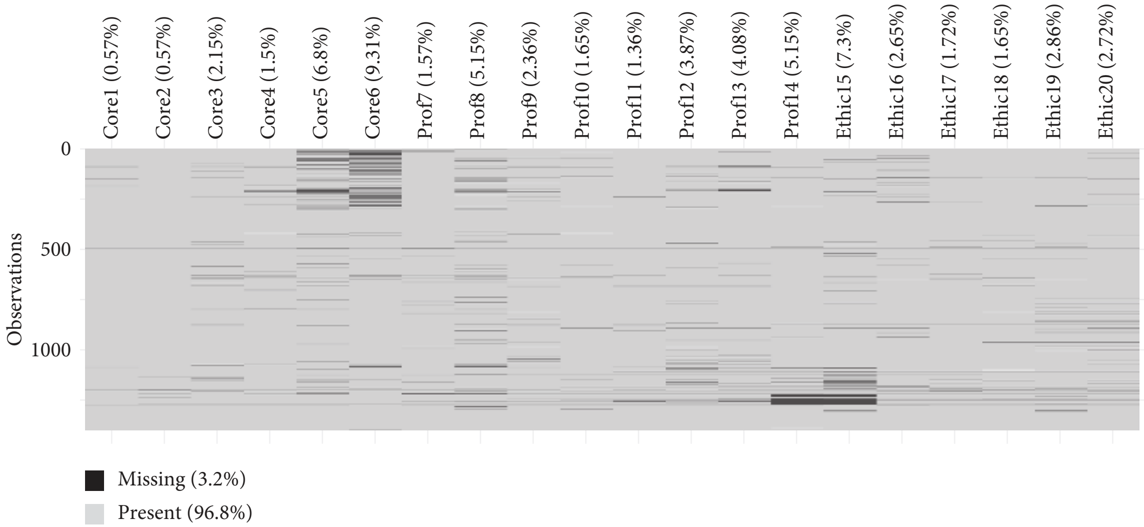


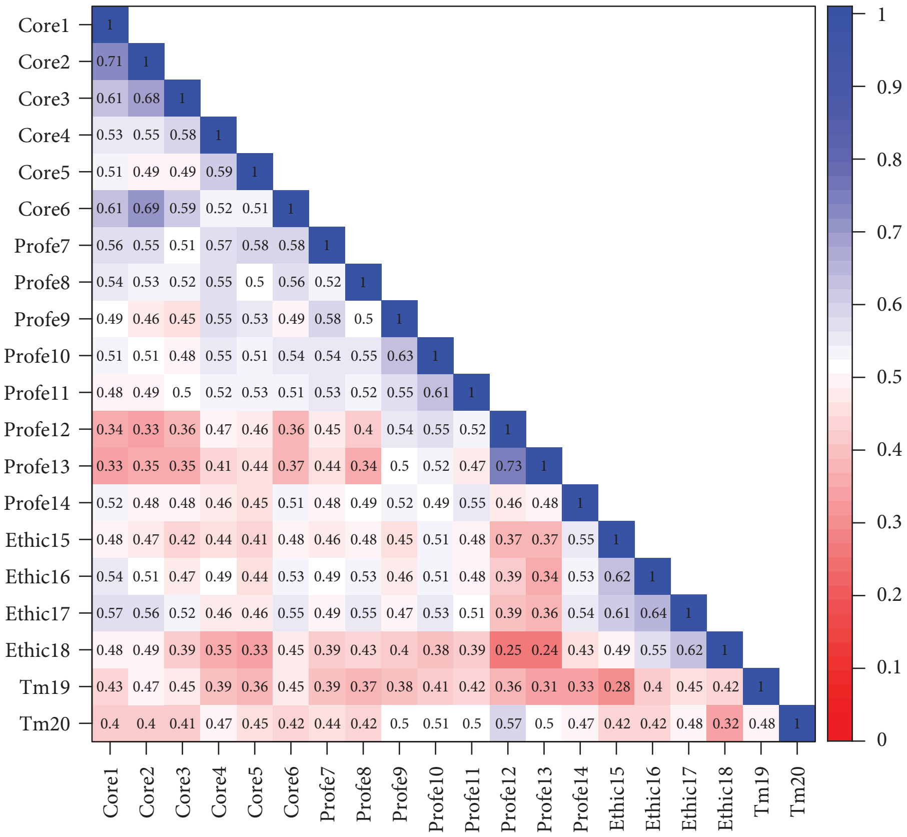
Figure 1: Heatmap visualization of missing data for the SETE data.

Table 1: Student evaluations of teacher's effectiveness scale.

Competency subscale (factor	Code	Measurement item used	VL	L	M	H	VH	NA
Core competency	Core1	Explains the course overall objectives, prepares course outline on time, and explains the contents of the course outline	1	2	3	4	5	
	Core2	Prepares well for course delivery	1	2	3	4	5	
	Core3	Gives course reading materials and lecture notes	1	2	3	4	5	
	Core4	Notifies list of references and textbooks available in the library	1	2	3	4	5	
	Core5	Teaches depending on course nature and teaches practical sessions	1	2	3	4	5	
	Core6	Delivers the course in a such a way that students understand	1	2	3	4	5	
Professional competency	Profe7	Uses of additional teaching aids	1	2	3	4	5	
	Profe8	Answers questions raised in the class room	1	2	3	4	5	
	Profe9	Gives class work, quiz, homework	1	2	3	4	5	
	Profe10	Uses student-centered approaches such as cooperative army, group work/ presentations.	1	2	3	4	5	
	Profe11	Follows continuous assessment approach and gives feedback on continuous assessments on time	1	2	3	4	5	
	Profe12	Gives supplementary exam to low-performing students on the basis of continuous assessment result	1	2	3	4	5	
	Profe13	Gives tutorial for female students, special needs students, and low-performing students	1	2	3	4	5	
	Profe14	Prepares exams as per the course content, exams cover across the course contents, exams include various assessment modes and allocates appropriate marks for exam questions	1	2	3	4	5	
Ethical competence	Ethic15	Gives respect to students	1	2	3	4	5	
	Ethic16	Listens to students' questions and gives feedback and allows students to interact during class room sessions	1	2	3	4	5	
	Ethic17	Ethics, behavior, and commitment for knowledge transfer	1	2	3	4	5	
	Ethic18	Does not discriminate on the basis of ethnic, religion, or sex	1	2	3	4	5	
Time management	Tm19	Appears on time during class timetable and uses class time appropriately	1	2	3	4	5	
	Tm20	Informs consultation hour and solves students' academic problems on time	1	2	3	4	5	

Source: The Federal Democratic Republic of Ethiopia Ministry of Science and Higher Education (May 29/2018), Ethiopia.

Discriminant validity was also checked by comparing the loading of an item across different constructs. If all items loaded more highly on the construct that they were measuring than on any other construct in the model, discriminant validity was met [57]. According to the EFA output presented in Table 4, considerable cross-loadings were observed. The nomological validity of the scale was checked by examining the significance of the construct correlation value between construct (interconstruct) variables in the model [35]. Accordingly, the 95%



	Core1	Core2	Core3	Core4	Core5	Core6	Profe7	Profe8	Profe9	Profe10	Profe11	Profe12	Profe13	Profe14	Ethic15	Ethic16	Ethic17	Ethic18	Tm19	Tm20
Core1	1																			
Core2	0.71	1																		
Core3	0.61	0.68	1																	
Core4	0.53	0.55	0.58	1																
Core5	0.51	0.49	0.49	0.59	1															
Core6	0.61	0.69	0.59	0.52	0.51	1														
Profe7	0.56	0.55	0.51	0.57	0.58	0.58	1													
Profe8	0.54	0.53	0.52	0.55	0.5	0.56	0.52	1												
Profe9	0.49	0.46	0.45	0.55	0.53	0.49	0.58	0.5	1											
Profe10	0.51	0.51	0.48	0.55	0.51	0.54	0.54	0.55	0.63	1										
Profe11	0.48	0.49	0.5	0.52	0.53	0.51	0.53	0.52	0.55	0.61	1									
Profe12	0.34	0.33	0.36	0.47	0.46	0.36	0.45	0.4	0.54	0.55	0.52	1								
Profe13	0.33	0.35	0.35	0.41	0.44	0.37	0.44	0.34	0.5	0.52	0.47	0.73	1							
Profe14	0.52	0.48	0.48	0.46	0.45	0.51	0.48	0.49	0.52	0.49	0.55	0.46	0.48	1						
Ethic15	0.48	0.47	0.42	0.44	0.41	0.48	0.46	0.48	0.45	0.51	0.48	0.37	0.37	0.55	1					
Ethic16	0.54	0.51	0.47	0.49	0.44	0.53	0.49	0.53	0.46	0.51	0.48	0.39	0.34	0.53	0.62	1				
Ethic17	0.57	0.56	0.52	0.46	0.46	0.55	0.49	0.55	0.47	0.53	0.51	0.39	0.36	0.54	0.61	0.64	1			
Ethic18	0.48	0.49	0.39	0.35	0.33	0.45	0.39	0.43	0.4	0.38	0.39	0.25	0.24	0.43	0.49	0.55	0.62	1		
Tm19	0.43	0.47	0.45	0.39	0.36	0.45	0.39	0.37	0.38	0.41	0.42	0.36	0.31	0.33	0.28	0.4	0.45	0.42	1	
Tm20	0.4	0.4	0.41	0.47	0.45	0.42	0.44	0.42	0.5	0.51	0.5	0.57	0.5	0.47	0.42	0.42	0.48	0.32	0.48	1

Figure 2: Interitem correlation matrix for SETE scale: The shades of the colors give an idea about the strength of the correlations. A strong positive correlation is indicated by blue color, while a weak positive correlation is indicated by red color.

Table 2: Reliability analysis of items (Cronbach's alpha) and the importance of each item for each construct ($n = 698$).

Factors and items	α value if an item is dropped	Item-total correlations	Corrected item-total correlation	M (SD)	Skewness
Core competency (α = 0.88)					
Explains the course's overall objectives, prepares the course outline on time, and explains the contents of the course outline	0.86	0.82	0.79	4.55 (0.85)	−2.28
Prepares well for course delivery	0.86	0.85	0.85	(0.84)	−2.19
Gives course reading materials and lecture notes	0.86	0.81	0.77	4.51 (0.87)	−2.06
Notifies list of references and textbooks available in the library	0.87	0.79	0.71	4.29 (1.04)	−1.63
Teaches depending on course nature teaches practical sessions	0.88	0.75	0.64	4.22 (1.1)	−1.57
Delivers the course in such a way that students understand	0.86	0.78	0.74	4.52 (0.85)	−2.20
Professional competency (α = 0.85)					
Uses of additional teaching aids	0.87	0.72	0.64	4.38 (0.98)	−1.80
Answers questions raised in the class room	0.88	0.64	0.55	4.51 (0.86)	−2.13
Gives class work, quiz, homework	0.87	0.78	0.7	4.25 (1.09)	−1.56
Uses a student-centered approach such as cooperative army, group work/presentations	0.87	0.8	0.72	4.28 (1.05)	−1.55
Follows continuous assessment approach and gives feedback on continuous assessments on time	0.87	0.76	0.68	4.35 (1)	−1.69
Gives supplementary exam to low-performing students on the basis of continuous assessment result	0.87	0.78	0.66	3.71 (1.46)	−0.84
Gives tutorial for female students, special needs students, and low-performing students	0.87	0.76	0.65	3.86 (1.39)	−1.02

Table 2: Continued.

Factors and items	α value if an item is dropped	Item-total correlations	Corrected item-total correlation	*M* (SD)	Skewness
Prepares exams as per the course content, exams cover across the course contents, exams include various assessment modes and allocates appropriate marks for exam questions	0.87	0.68	0.59	4.45 (0.94)	−2.04
Ethical competence (α = 0.84)					
Gives respect to students	0.82	0.83	0.66	4.56 (0.86)	−2.42
Listens to students' questions and gives feedback and allows students to interact during class room sessions	0.79	0.83	0.71	4.62 (0.78)	−2.41
Ethics, behavior, and commitment for knowledge transfer	0.78	0.86	0.73	4.59 (0.85)	−2.47
Does not discriminate on the basis of ethnicity, religion, or sex	0.83	0.82	0.7	4.72 (0.7)	−3.18
Time management (α = 0.64)					
Appears on time during class timetable and uses class time appropriately	0.48	0.8	0.44	4.56 (0.87)	−2.28
Informs consultation hour and solves students' academic problems on time	0.23	0.89	0.44	4.27 (1.14)	−2.19

Table 3: Validation measures for the four-factor model.

	CR	AVE	ASV	MSV	CC	PC	EC	TM
CC	0.88	0.56	0.68	0.76	1			
PC	0.88	0.5	0.71	0.76	0.98	1		
EC	0.85	0.6	0.59	0.64	0.86	0.87	1	
TM	0.67	0.51	0.64	0.76	0.87	0.86	0.91	1

The highlighted nondiagonal numbers are Heterotrait-monotrait (HTMT) ratios of the correlation between the two constructs.

Table 4: Factor loadings and the total variance explained for the factors in explanatory factor analysis (EFA).

Factors and items	Factor 1	Factor 2	h^2	U
Core competency				
Explains the course overall objectives, prepares the course outline on time, and explains the contents of the course outline	**0.65**	0.36	0.54	0.46
Prepares well for course delivery	**0.68**	0.29	0.55	0.45
Gives course reading materials and lecture notes	**0.66**	0.36	0.57	0.43
Notifies list of references and textbooks available in the library	**0.52**	0.52	0.54	0.46
Teaches practical sessions depending on the course nature	0.31	***0.56***	0.48	0.52
Delivers the course in a such a way that students understand	**0.66**	0.28	0.51	0.49
Professional competency				
Uses of additional teaching aids	**0.55**	0.55	0.56	0.44
Answers questions raised in the class room	**0.61**	0.38	0.51	0.49
Gives class work, quiz, homework	0.42	***0.63***	0.58	0.42
Uses student-centered approach such, group work/presentations, etc.	0.36	***0.61***	0.58	0.42
Follows continuous assessment approach and gives feedback on continuous assessments on time	0.37	***0.59***	0.57	0.43
Gives supplementary exam to low-performing students on the basis of continuous assessment result	0.18	***0.85***	0.75	0.25
Gives tutorial for female students, special needs students, and low-performing students	0.14	***0.84***	0.72	0.28
Prepares exams as per the course content, exams cover across the course contents, exams include various assessment modes and allocates appropriate marks for exam questions	**0.59**	0.35	0.56	0.44
Ethical competence				
Gives respect to students	**0.64**	0.29	0.49	0.51
Listens to students' questions and gives feedback and allows students to interact during class room sessions	**0.71**	0.26	0.58	0.42
Ethics, behavior, and commitment for knowledge transfer	**0.7**	0.33	0.6	0.4
Does not discriminate on the basis of ethnic, religion, or sex	**0.74**	0.06	0.55	0.45
Time management				
Appears on time during class timetable and uses class time appropriately	**0.48**	0.36	0.36	0.64
Informs consultation hour and solves students' academic problems on time	0.31	**0.65**	0.52	0.48
Rotation sums of squared loadings				
% of variance	31	25		
Cumulative%	31	56		
Cronbach's alpha (computed with "testing" data)	0.93	**0.87**		

Table 5: Factor Loadings of the indicator items for the two-factor model.

	Beta	P
Factor 1		
Explains the course overall objectives, prepares the course outline on time, and explains the contents of the course outline	0.62	<0.001
Prepares well for course delivery	0.61	<0.001
Gives course reading materials and lecture notes	0.60	<0.001
Delivers the course in such a way that students understand	0.61	<0.001
Answers questions raised in the class room	0.63	<0.001
Prepares exams as per the course content, exams cover across the course contents, exams include various assessment modes and allocates appropriate marks for exam questions	0.70	<0.001
Listens to students' questions and gives feedback and allows students to interact during class room	0.62	<0.001
Ethics, behavior, and commitment for knowledge transfer	0.60	<0.001
Does not discriminate on the basis of ethnic, religion, or sex	0.68	<0.001
Listens to students' questions and gives feedback and allows students to interact during class room sessions	0.41	<0.001
Appears on time during class timetable and uses class time appropriately	0.49	<0.001
Factor 2		
Teaches depending on course nature and teaches practical sessions	0.78	<0.001
Gives class work, quiz, homework	0.81	<0.001
Uses student-centered approach such, group work/presentations	0.90	<0.001
Follows continuous assessment approach and gives feedback on continuous assessments on time	0.80	<0.001
Gives supplementary exam to low-performing students on the basis of continuous assessment result	1.00	<0.001
Gives tutorial for female students, special needs students, and low-performing students	0.90	<0.001
Informs consultation hour and solves students' academic problems on time	0.83	<0.001

confidence interval for interconstruct correlations in Table 3 does not contain 1, implying a statistically significant interconstruct correlation. This shows the poor nomological validity of the SETE scale.

3.4. The Factor Structure of the Scale.

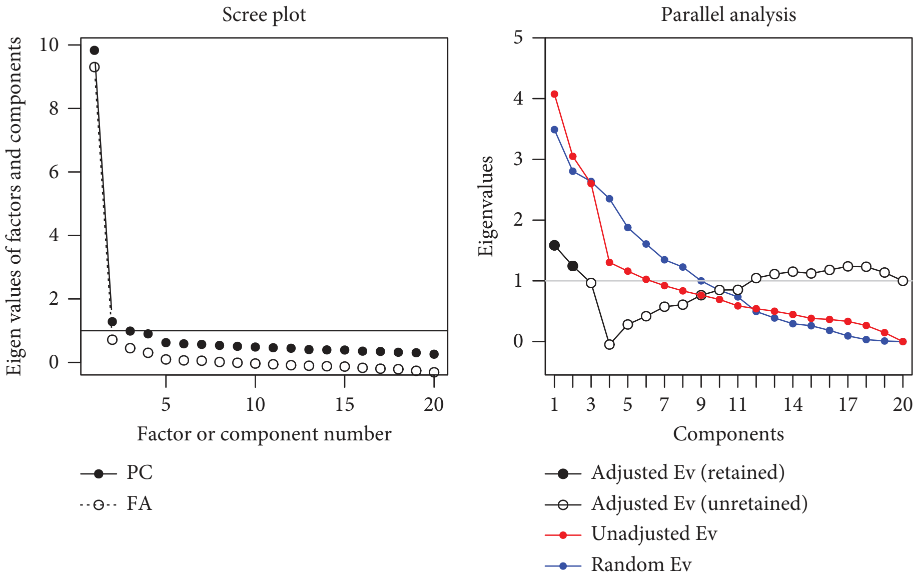
In this analysis, we used twofold cross-validation (CV) such that the data were divided into two random samples. The first half of the dataset with 699 observations (called the training data) was used to find the possible factor structure of the SETE scale using exploratory factor analysis, and the second half having 698 observations (called the testing data) was used to verify the factor structure of the scale.

3.4.1. Exploratory Factor Analysis.

An exploratory factor analysis (EFA) using the varimax-rotated component method was performed on the training data to check if item grouping was consistent with the proposed theory, that is, to test the structural validity. Before conducting factor analysis, the item-to-item correlation was examined by conducting the Kaiser–Meyer–Olkin (KMO) test and Bartlett's test for sphericity to see if there is a certain redundancy between the variables that we can summarize with the factors. The value of KMO was 0.96 and Bartlett's test of Sphericity produced $p < 0.001$, which are wonderful values. Thus, all variables could be considered for EFA [45,58].

We applied the scree plot test [59] and parallel analysis [60] to determine the required number of factors to retain. The rule for scree plots is to retain the factors above the point where the curve starts to level off (inflection point) and eliminate any factor below the inflection point [61]. From the scree plot (left panel in Figure 3), the first two factors of the scale have eigen values greater than one. Parallel analysis offers a more objective way to assess the appropriate number of components, where factors with adjusted eigenvalues greater than one are retained [62]. Both methods suggest retaining two factors.

The preliminary exploratory factor analysis (EFA) results, described in Table 6, revealed that the item variables are not significantly grouped under the respective factors, as theoretically defined. Hence, the factor structure of the scale is not consistent with the proposed understanding of the intention of the experts who devised the scale, indicating that the results from the EFA did not support the theoretical factor structure. The h^2 column in the table represents the value of communality, which must be higher than 0.3. The root mean square of the residuals (RMSR) was 0.05. Additionally, the root mean square of the residuals (RMSR = 0.05) is less than 0.1, verifying that the retained factors are appropriate for describing the correlation structure. From the results presented in Table 6, all items demonstrated high loading, ranging from 0.48 to 0.85, implying that all items are considered as important. Items in italics are loaded in the second factor. The factor loading of the first factor ranged from 0.48 to 0.77, while the factor loadings of the second factor ranged from 0.59 to 0.85. The analysis output includes the explained variance ratio. The first factor explained 31% of the total variance. The second factor explained 25% of the total variance. Hence, the two-factor construct explains 56% of the total variance. The analysis output includes the interfactor correlation after the explained variance ratio section. Based on our qualitative judgment of item content, the less serious nature of cross loading, and the expected association of factors, we decided to keep these items and assign them to the factor in which they showed stronger factor loadings and were found the most relevant. However, the two items "core 5" and "profe7" load equally on the two items. Hence, we decided to remove them.



(a)

(b)

Figure 3: Parallel analysis and scree plots for the 20-item SETE scale. PC: Principal components, FA: Factor analysis.

Table 6: Interconstruct correlations and their 95% CI.

Constructs	√AVE	CC [95% CI]	PC [95% CI]	EC [95% CI]	TM [95% CI]
CC	0.75	1			
PrC	0.71	0.87 [0.84; 0.90]	1		
EC	0.75	0.81 [0.75; 0.86]	0.78 [0.74; 0.83]	1	
TM	0.71	0.8 [0.7; 0.89]	0.87 [0.77; 0.96]	0.73 [0.63; 0.83]	1

3.4.2. Results of CFA for the Underlying Structure. The models analyzed were identified, which means that there should be more observations than the parameters to be estimated [63].

Confirmatory factor analysis was carried out to check if the number of factors (or constructs) and the indicator variables conformed to what was expected based on the theory. Multiple fit indices were used to evaluate whether the models adequately reflected the observed data. Moreover, the two models were compared to assess if they had an identical fit. We used the "Testing" data for this purpose.

Figure 4 presents the path diagram of the confirmatory factor analysis for two-factor (left panel) and four-factor (right panel) models, where a single-headed arrow is used to imply a direction of the assumed causal influence, and double-headed arrows are used to represent the covariance between two latent variables (factors).

From the path diagram for the two-factor model, the measurement error ranged between 0.36 (Profe10) and 0.61 (Profe13 and Profe13). Similarly, the four-factor model produced a measurement error ranging between 0.30 (Ethic17) and 0.59 (TM19). The increase in the measurement error for the two-factor model is due to specifying a relatively less number of factors than expected [43].

For the two-factor model, it was thus deduced that the squared coefficient of multiple correlations or the amount of variance explained by the latent variable fell within a range between 0.75 and 0.48. Similarly, all factor loadings had values equal to or greater than 0.61 (p13). The correlations between latent constructs ranged between 0.73 and 0.87. The interconstruct correlations of core competency with professional competency, ethical quality, and time management were 0.87, 0.8, and 0.7, respectively. Similarly, interconstruct correlations of professional competency with ethical quality were 0.78 and 0.87, whereas interconstruct correlation between ethical quality and time management was 0.73.

Table 5 shows that the standardized coefficients for the two-factor model are significant at the 0.001 level, implying that all items are significantly correlated with their respective constructs. Because the domain is standardized (mean = 0, SD = 1), the coefficients are interpreted as the increase (or decrease) in the score of an item for every standard deviation increase in the factor/construct. For example, $\beta = 0.69$, that is, for every standard deviation increase in core competency, "Core1" increases by 0.69. In addition, in the SETE scale, the " profe12" item had the highest association with its construct ($\beta = 1$). The values in Table 7 can be interpreted similarly.

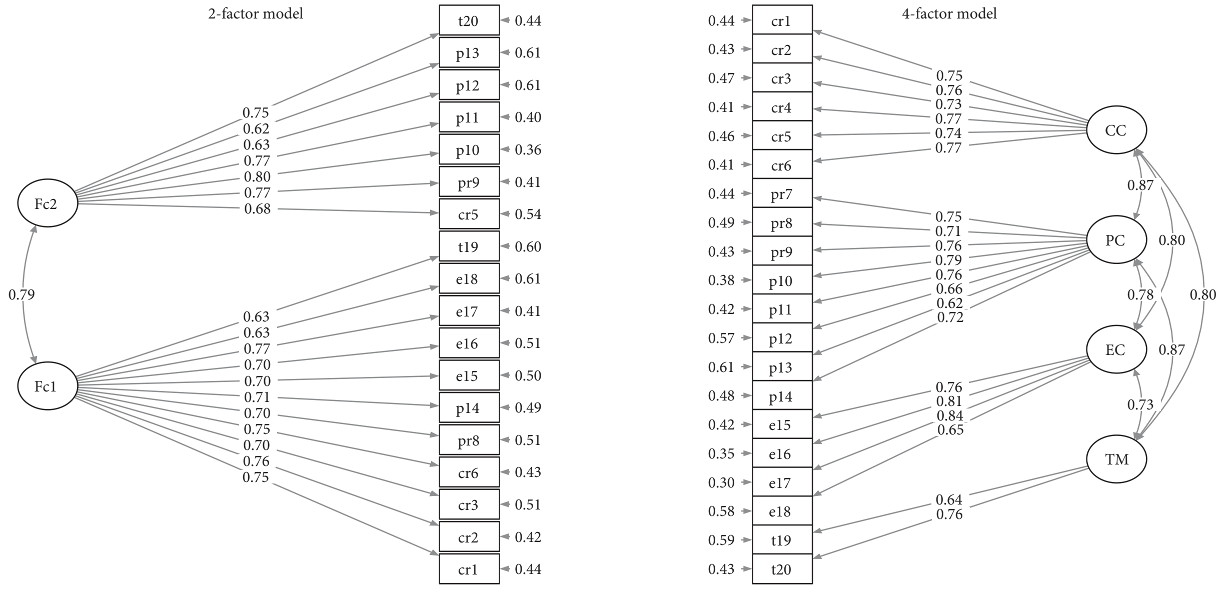


FIGURE 4: Path Diagram for the two Confirmatory Factor Models. Fc1: factor 1; Fct2: factor 2; CC: Core competency; PC: Professional competency; EC: Ethical competency; TM: Time management.

TABLE 7: Factor Loadings of the indicator items for the four-factor model.

Factors and items	Beta	P
Core competency		
Explains the course overall objectives, prepares the course outline on time, and explains the contents of the course outline	0.67	<0.001
Prepares well for course delivery	0.70	<0.001
Gives course reading materials and lecture notes	0.65	<0.001
Notifies the list of references and textbooks available in the library	0.69	<0.001
Teaches depending on course nature and teaches practical sessions	0.68	<0.001
Delivers the course in a such a way that students understand	0.59	<0.001
Professional competency		
Uses additional teaching aids	0.66	<0.001
Answers questions raised in the class room	0.52	<0.001
Gives class work, quiz, homework	0.77	<0.001
Uses student-centered approach, such as group work/presentations	0.77	<0.001
Follows continuous assessment approach and gives feedback on continuous assessments on time	0.70	<0.001
Gives supplementary exam to low-performing students on the basis of continuous assessment result	0.91	<0.001
Gives tutorial for female students, special needs students, and low-performing students	0.84	<0.001
Prepares exams as per the course content, exams cover across the course contents, exams include various assessment modes and allocates appropriate marks for exam questions	0.55	<0.001
Ethical competence		
Gives respect to students	0.61	<0.001
Listens to students' questions and gives feedback and allows students to interact during class room sessions	0.55	<0.001
Ethics, behavior, and commitment for knowledge transfer	0.66	<0.001
Does not discriminate on the basis of ethnic, religion, or sex	0.50	<0.001
Time management		
Appears on time during class timetable and use class time appropriately	0.52	<0.001
Informs consultation hour and solves students' academic problems on time	0.80	<0.001

3.4.3. Model Fit and Comparison. The appropriateness of the measurement model in comparison with the data was examined first. The best model should have a relative chi-square (χ^2/df) value close to 1.

We also used the comparative fit index (CFI) and Tucker–Lewis index (TLI) and RMSEA to measure whether the model fits the data better than a more restricted baseline model. However, the cutoff values for these indices are arbitrary, and the meaning of "good" fit and its relationship with fit indices are not well understood [64].

The absolute and comparative fit indices for the two-factor and four-factor CFA models are presented in Table 8. The comparative fit parameters for the four-factor model, CFI (0.89) and TLI (0.87) are less than the acceptable cutoff point of 0.90, which is relatively poor fit [65]. The comparative fit indices of the four-factor model are 0.088

Table 8: Absolute and comparative fit indices for the two- and four-factor CFA models.

Fit indices	Two-factor model	Four-factor model
Absolute fit index		
Relative χ^2	1.03	1.2
Root mean square error of approximation (RMSEA)	0.008	0.088
Standardized root mean square residual(SRMR)	0.056	0.06
p-value RMSEA ≤ 0.05	1	<0.00001
Incremental fit index		
Comparative fit index (CFI)	0.999	0.89
Tucker–Lewis (TLI)	0.999	0.87

(RMSEA) and 0.06 (SRMR), which are considered an indication of fair fit [66].

However, for the two-factor model, the comparative fit indices, CFI (0.999), and TLI (0.999) were greater than the 0.90 threshold, indicating an improvement of the tested four-factor model in a relative sense. We also found that the SRMR (0.056) had a good fit (<0.06), and RMSEA (0.008) had a good fit (<0.05), indicating that the two-factor model fits well to the data [67].

In conclusion, the two-dimensional model provided improved goodness-of fit indices than the four-factor model, implying that the two-factor model fits the data better than the four-factor model.

Test of comparison of the two models to explain the factor structure of the scale showed a nonsignificant p value, showing that the four-factor model did not do a better job than the two-factor model. Moreover, for the AIC, a value of 29661.43 was obtained for the two-factor model and a value of 29677.60 for the four-factor model. Thus, the two-factor model should be preferred (smaller AIC).

4. Discussion

In educational institutions, evaluating teachers' effectiveness is similar to evaluating students' learning [31]. Student evaluations of teachers' effectiveness are a current and controversial topic in higher education and research. Many stakeholders, including teachers, are doubtful of SETE's effectiveness and validity for both formative and summative purposes [7,68]. Thus, the primary goal of this study was to look into the psychometric properties of the students' assessment of the SETE scale, which is used by Ethiopian higher education institutions.

From the results, the SETE scale was shown to have good internal consistency and good convergent validity. This result complements the findings of [7,19–30, 69], although the dimensionality and number of items of these scales are unrelated. However, unlike the student evaluation of higher education teachers' effectiveness scale developed by [18,19,21,22,25,31,70], the SETE scale used by Ethiopian higher education faced a validity problem. Moreover, the CFA results showed poor fit indices, revealing that the underlying four-factor structure for the SETE scale is insufficient to explain the data structure. This is because the SETE scale was developed based on the evaluation on theoretical grounds. However, its development should have gone through quantitative exploration in addition to the experts' evaluation on theoretical grounds, which is one of the criteria to ensure content validity. Scale development is not a straightforward endeavor [71]. Hinkin [72] pointed out three phases of scale development to create a rigorous scale: item development (consisting of steps of identification of the domains, item generation, and content validity or theoretical analysis), scale development (including steps pretesting the items in the scale, survey administration and sample size, item reduction analysis, extraction of factors), and scale evaluation (consisting of tests of dimensionality, reliability, and validity). According to researchers' ample experience during the development of the SETE scale, however, its development fails to follow the procedures used by Hinkin [72].

5. Conclusion

The construction of valid and reliable scales requires systematic research, in which both theoretical knowledge and empirical data should play an important role. This study is the first attempt to assess the validity of the SETE scale, which is used by Ethiopian higher education institutions. The current study attempted to provide evidence of convergent validity, discriminant validity, and nomological validity of the SETE scale that Ethiopian public higher education institutions used to evaluate their teachers' performance. Accordingly, the scale lacks both discriminant and nomological validity despite its convergent validity, revealing that the SETE scale does not appear to discriminate well among the constructs it measures.

Although further research is needed to confirm these results based on multicenter data, the two-factor model with 18 items yielded a better factor structure of the SETE scale. This is because the dimensionality of the scale was developed based on the opinion of experts only; it did not necessarily measure the important competency components of the teachers. Overall, the findings indicate that the SETE scale cannot be used to effectively assess teachers' teaching effectiveness unless further improvements are made to the scale and its development process.

This work has practical, theoretical, and policy implications for a variety of stakeholders at various levels. In practice, this research can assist higher education institutions and the Ministry of Science and Higher Education in identifying the SETE scale's psychometric gaps. As a result, it can be used as a framework for improving the instrument's reliability and validity in order to clearly measure teachers' effectiveness and, as a result, propose interventions to increase teachers' performance and motivation. The findings of this study can also be used to offer new knowledge and concepts on the assessment of teachers' performance and pedagogical competencies in higher education, especially in Ethiopia. As previously stated, no investigation on the psychometric features of the SETE scale has been done in Ethiopia.

6. Research Limitations and Future Directions

The results of this study should be considered in the light of these limitations. One limitation was that although the scale is harmonized and used by all public universities, this analysis used data from a single university, which may not be generalizable to the remaining public universities across the country. Hence, this study emphasizes the need to obtain large amounts of data from multiple universities to further strengthen the outcomes of the study. The study also assumed that students rated their teachers with no bias or prejudice. However, it is well perceived from experience that students who receive higher grades in the course rate teachers more favorably, whereas low-grade achievers revenge their teachers in the form of low teacher ratings. Other factors such as time of evaluation, physical attractiveness of the teacher, course difficulty, age, and the teacher's personality influence student ratings [28,73]. Despite its convenience, the current study used one dataset for both PCA and CFA; hence, further studies are needed to validate both the SETE scale framework and measures. Careful planning of the validation process should be carried out with large data to obtain stronger evidence on the findings and develop a scale that measures teaching effectiveness appropriately. Furthermore, analysis at a different point in time needs to be carried out to test the test-retest reliability of the scale. Although the maximum number of items per scale will depend on the complexity of the variable being measured, increasing the number of items per scale improves the scale's richness to capture more information [74]. However, the "Time management" subscale has only two items, which is another limitation of this study.

Abbreviations

CFA:	Confirmatory factor analysis
CR:	Composite reliability
EFA:	Exploratory factor analysis
Ev:	Eigen value
HTMT:	Heterotrait-monotrait
PCA:	Principal component analysis
SETE:	Student evaluation of teaching effectiveness.

Appendix

Student Evaluations of Teacher's Effectiveness Scale

Dear student. Please check in the boxes indicating how you evaluate your teachers for this semester altogether.

Date: _______ Your field of study: _______Year: ______
Your gender: ______ Course: _______

Based on the evaluation point, rate by placing a circle on any of the ranks indicated, ranging from very low to very high. Note: VL = very low; L = low; M = medium = M; H = high; VH = very high; NA = not applicable.

Data Availability

The datasets analyzed in this study are available from the corresponding author on reasonable request.

Ethical Approval

The researchers have got permission from Debre Markos University Quality Assurance Office, to use the data without fabrication and falsification. As the study was based on secondary data, informed consent was not obtained from the study participants, but the anonymity and the confidentiality of the data were assured.

Conflicts of Interest

The authors declare that none of them has a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper.

Authors' Contributions

MGA contributed to the study concept and design of the statistical methodology, performed the analysis and interpretation of the data, and wrote the first draft of the manuscript. AAD and DMF contributed to the study by critically revising the manuscript. All the authors read and approved the final manuscript.

Acknowledgments

The authors are grateful to the Debre Markos University Quality Assurance Office for permission to use the data. This work was financially supported by Debre Markos University.

Supplementary Materials

R codes used for assessing scale validity (doc 354 KB) are included. (*Supplementary Materials*)

References

[1] O. Bolarinwa, "Principles and methods of validity and reliability testing of questionnaires used in social and health science researches," *Nigerian Postgraduate Medical Journal*, vol. 22, no. 4, p. 195, 2015.

[2] A. T. Ginty, "Psychometric properties," in *Encyclopedia of Behavioral Medicine*, M. D. Gellman and J. R. Turner, Eds., Springer New York, New York, NY, pp. 1563-1564, 2013.

[3] N. Asiamah, K. Kouveliotis, R. Eduafo, and R. Borkey, "Psychometric properties of a new scale measuring neglect and abuse of older adults in the community: implications for

social activity," *International Quarterly of Community Health Education*, vol. 41, no. 2, pp. 163–172, 2021.

[4] F. E. Espinoza Molina, B. d. V. Arenas Ramirez, F. Aparicio Izquierdo, and D. C. Zúñiga Ortega, "Road safety perception questionnaire (RSPQ) in Latin America: a development and validation study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 5, p. 2433, 2021.

[5] H. K. Mohajan, "Two criteria for good measurements in research: validity and reliability," *Annals of Spiru Haret University. Economic Series*, vol. 17, no. 4, pp. 59–82, 2017.

[6] K. De Witte and N. Rogge, "Accounting for exogenous influences in performance evaluations of teachers," *Economics of Education Review*, vol. 30, no. 4, pp. 641–653, 2011.

[7] P. Spooren, B. Brockx, and D. Mortelmans, "On the validity of student evaluation of teaching," *Review of Educational Research*, vol. 83, no. 4, pp. 598–642, 2013.

[8] S. Embiza and S. Hadush, "Assessing the dimensionality and reliability of teachers' performance evaluation in eastern zone high schools, tigrai national regional state, Ethiopia," *Journal of Education and Practice*, vol. 6, no. 7, pp. 91–99, 2015.

[9] D. N. Harris and T. R. Sass, "Teacher training, teacher quality and student achievement," *Journal of Public Economics*, vol. 95, no. 7-8, pp. 798–812, 2011.

[10] J. Kagema and C. Irungu, "An analysis of teacher performance appraisals and their influence on teacher performance in secondary schools in Kenya," *International Journal of Education*, vol. 11, no. 1, pp. 93–98, 2018.

[11] X.-f. Zhang and H.-m. Ng, "An effective model of teacher appraisal," *Educational Management Administration & Leadership*, vol. 45, no. 2, pp. 196–218, 2017.

[12] M. Getie Ayaneh, A. A. Dessie, and A. W. Ayele, "Survival models for the analysis of waiting time to first employment of new graduates: a case of 2018 Debre Markos university graduates, northwest Ethiopia," *Education Research International*, vol. 2020, Article ID 8877504, 10 pages, 2020.

[13] T. Sánchez, R. Gilar, J.-L. Castejon, J. Vidal, and J. León, "Students' evaluation of teaching and their academic achievement in a higher education institution of Ecuador," *Frontiers in Psychology*, vol. 11, 2020.

[14] M. D. Toland and R. J. De Ayala, "A multilevel factor Analysis of students' evaluations of teaching," *Educational and Psychological Measurement*, vol. 65, no. 2, pp. 272–296, 2005.

[15] P. Ginns, M. Prosser, and S. Barrie, "Students' perceptions of teaching quality in higher education: the perspective of currently enrolled students," *Studies in Higher Education*, vol. 32, no. 5, pp. 603–615, 2007.

[16] D. Mortelmans and P. Spooren, "A revalidation of the SET37 questionnaire for student evaluations of teaching," *Educational Studies*, vol. 35, no. 5, pp. 547–552, 2009.

[17] D. C. Barnes, B. Engelland, C. Matherne, W. Martin, C. Oregon, and K. Ring, "Developing a psychometrically sound measure of collegiate teaching proficiency," *College Student Journal*, vol. 42, no. 1, pp. 199–214, 2008.

[18] I. Altaf, A. Kamal, and B. Hassan, "Development and validation of university teacher's evaluation scale," *Pakistan Journal of Psychological Research*, vol. 28, no. 1, pp. 155–178, 2013.

[19] S. Shahzad and N. Mahmood, "Development of teaching effectiveness scale for university teachers," *Journal of Social Science Research*, vol. 7, no. 2, 2019.

[20] P. Barman, D. Bhattacharyya, and P. Barman, "Teaching effectiveness of teacher educators in different types of B. Ed colleges in West Bengal, India," *American Journal of Educational Research*, vol. 3, no. 11, pp. 1164–1177, 2015.

[21] İ. Marulcu and K. Bozkuş, "Adaptation of the teacher effectiveness scale in higher education into Turkish language," *European Journal of Education Studies*, vol. 3, no. 10, 2017.

[22] S. Mittal, R. Gera, and D. K. Batra, "Evaluating the validity of student evaluation of teaching effectiveness (SET) in India," *Education+Training*, vol. 57, 2015.

[23] D. Feistauer and T. Richter, "How reliable are students' evaluations of teaching quality? A variance components approach," *Assessment & Evaluation in Higher Education*, vol. 42, no. 8, pp. 1263–1279, 2017.

[24] H. W. Marsh and M. Bailey, "Multidimensional students' evaluations of teaching effectiveness," *The Journal of Higher Education*, vol. 64, no. 1, pp. 1–18, 1993.

[25] P.-T. Oon, B. Spencer, and C. C. S. Kam, "Psychometric quality of a student evaluation of teaching survey in higher education," *Assessment & Evaluation in Higher Education*, vol. 42, no. 5, pp. 788–800, 2017.

[26] K. Otani, B. J. Kim, and J.-I. Cho, "Student evaluation of teaching (SET) in higher education: how to use SET more effectively and efficiently in public affairs education," *Journal of Public Affairs Education*, vol. 18, no. 3, pp. 531–544, 2012.

[27] D. E. Clayson, "Student evaluation of teaching and matters of reliability," *Assessment & Evaluation in Higher Education*, vol. 43, no. 4, pp. 666–681, 2018.

[28] M. Kelly, *Student Evaluations of Teaching Effectiveness: Considerations for Ontario Universities*, Council of Ontario Universities, Toronto, Canada, 2012.

[29] T. L. Khong, "The validity and reliability of the student evaluation of teaching: a case in a private higher educational institution in Malaysia," *International Journal for Innovation Education and Research*, vol. 2, no. 9, pp. 57–63, 2016.

[30] M. Shevlin, P. Banyard, M. Davies, and M. Griffiths, "The validity of student evaluation of teaching in higher education: love me, love my lectures?" *Assessment & Evaluation in Higher Education*, vol. 25, no. 4, pp. 397–405, 2000.

[31] G. M. Calaguas, "Teacher effectiveness scale in higher education: development and psychometric properties," *International Journal of Research Studies in Education*, vol. 1, no. 1, pp. 1–18, 2012.

[32] MOE, Ethiopian Education Development Roadmap, 2018.

[33] MOE, Higher Education: Performance Indicator Reference Sheet. 2017.

[34] D. D. Sozo and M. T. Kabtyimer, "Evaluation of teacher-s'competency in higher education: the case of evaluation by students in Arba Minch University, Ethiopia," *European Journal of Education Studies*, vol. 7, no. 4, 2020.

[35] A. N. Adedeji, M. O. Lawan, and S. F. Sidique, "Testing validity of observed indicators of local content policy in Nigeria: evidence from four-factor measurement model," *CBN Journal of Applied Statistics*, vol. 8, no. 1, pp. 149–173, 2017.

[36] K. A. Engellant, D. D. Holland, and R. T. Piper, "Assessing convergent and discriminant validity of the motivation construct for the technology integration education (TIE) model," *Journal of Higher Education Theory & Practice*, vol. 16, no. 1, 2016.

[37] R. Aghekyan, "Measuring high school students' science identities, expectations of success in science, values of science and environmental attitudes: development and validation of the SIEVEA survey," *Science Education International*, vol. 30, no. 4, pp. 342–353, 2019.

[38] C. Fornell and D. F. Larcker, *Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics*, Sage Publications Sage CA, Los Angeles, CA, USA, 1981.

[39] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2015.

[40] C.-F. Wang, R.-H. Wang, P.-Y. Yen, S.-L. Huang, Y.-B. Huang, and Y.-C. Lin, "Construction and validation of a teacher self-evaluation scale to measure teaching performance in a medical university," *Journal of Medical Education*, vol. 24, no. 4, pp. 195–208, 2020.

[41] J. Haslbeck and R. van Bork, *Estimating the Number of Factors in Exploratory Factor Analysis via Out-Of-Sample Prediction Errors*, 2021, https://psyarxiv.com/.

[42] H. Todorov, D. Fournier, and S. Gerber, "Principal components analysis: theory and application to gene expression data analysis," *Genomics and Computational Biology*, vol. 4, no. 2, Article ID e100041, 2018.

[43] J. C. Hayton, D. G. Allen, and V. Scarpello, "Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis," *Organizational Research Methods*, vol. 7, no. 2, pp. 191–205, 2004.

[44] B. Thompson, *Exploratory and Confirmatory Factor Analysis*, American Psychological Association, Washington, DC, USA, 2004.

[45] R. P. Sarmento and V. Costa, "Confirmatory factor analysis--a case study," 2019, https://arxiv.org/abs/1905.05598.

[46] Y. Rosseel, Package "lavaan", 2017.

[47] R. C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020, http://www.R-project.org..

[48] N. Tierney, "Visdat: visualising whole data frames," *The Journal of Open Source Software*, vol. 2, no. 16, p. 355, 2017.

[49] C.-H. Li, "Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares," *Behavior Research Methods*, vol. 48, no. 3, pp. 936–949, 2016.

[50] D. Mindrila, "Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: a comparison of estimation bias with ordinal and multivariate non-normal data," *International Journal of Digital Society*, vol. 1, no. 1, pp. 60–66, 2010.

[51] C. DiStefano and G. B. Morgan, "A comparison of diagonal weighted least squares robust estimation techniques for ordinal data," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 21, no. 3, pp. 425–438, 2014.

[52] J. C. Nunnally, "An overview of psychological measurement," *Clinical Diagnosis of Mental Disorders*, pp. 97–146, 1978.

[53] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, Guilford Publications, USA, 2015.

[54] R. L. Piedmont, "Inter-item correlations," in *Encyclopedia of Quality of Life and Well-Being Research*, A. C. Michalos, Ed., Springer Netherlands, Dordrecht, Netherlands, pp. 3303-3304, 2014.

[55] J. Henseler, "Bridging design and behavioral research with variance-based structural equation modeling," *Journal of Advertising*, vol. 46, no. 1, pp. 178–192, 2017.

[56] K. A. Markus, *Principles and Practice of Structural Equation Modeling by Rex B. Kline*, Taylor & Francis, UK, 2012.

[57] D. Gefen, D. Straub, and M.-C. Boudreau, "Structural equation modeling and regression: guidelines for research practice," *Communications of the Association for Information Systems*, vol. 4, no. 1, p. 7, 2000.

[58] N. Manninen, E. Kuusisto, and K. Tirri, "Finnish social services students' perceptions of purpose and helping unknown others," *International Journal of Social Pedagogy*, vol. 8, no. 1, pp. 1–13, 2019.

[59] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966.

[60] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185, 1965.

[61] A. Onatski, "Determining the number of factors from empirical distribution of eigenvalues," *Review of Economics and Statistics*, vol. 92, no. 4, pp. 1004–1016, 2010.

[62] A. Dinno and M. A. Dinno, *Package "Paran"*, R Package Version, Dortmund, Germany, 2018.

[63] M. A. Verdugo, V. M. Guillén, B. Arias, E. Vicente, and M. Badia, "Confirmatory factor analysis of the supports intensity scale for children," *Research in Developmental Disabilities*, vol. 49-50, pp. 140–152, 2016.

[64] K. Lai and S. B. Green, "The problem with having two watches: assessment of fit when RMSEA and CFI disagree," *Multivariate Behavioral Research*, vol. 51, no. 2-3, pp. 220–239, 2016.

[65] K. Joreskog and D. Sorbom, *Structural Equation Modelling: Guidelines for Determining Model Fit*, University Press of America, Lanham, MD, USA, 1993.

[66] K. Schermelleh-Engel, H. Moosbrugger, and H. Müller, "Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures," *Methods of Psychological Research Online*, vol. 8, no. 2, pp. 23–74, 2003.

[67] S. Cangur and I. Ercan, "Comparison of model fit indices used in structural equation modeling under multivariate normality," *Journal of Modern Applied Statistical Methods*, vol. 14, no. 1, p. 14, 2015.

[68] L. P. Aultman, "An unexpected benefit of formative student evaluations," *College Teaching*, vol. 54, no. 3, pp. 251–285, 2006.

[69] M. H. Knol, C. V. Dolan, G. J. Mellenbergh, and H. L. J. van der Maas, "Measuring the quality of university lectures: development and validation of the instructional skills questionnaire (ISQ)," *PloS One*, vol. 11, no. 2, Article ID e0149163, 2016.

[70] H. W. Marsh, "Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness," in *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, pp. 319–383, Springer, Berlin, Germany, 2007.

[71] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young, "Best practices for developing and validating scales for health, social, and behavioral research: a primer," *Frontiers in Public Health*, vol. 6, p. 149, 2018.

[72] T. R. Hinkin, "A review of scale development practices in the study of organizations," *Journal of Management*, vol. 21, no. 5, pp. 967–988, 1995.

[73] J. W. Lawrence, *Student Evaluations of Teaching Are Not Valid*, American Association of University Professors, Washington, DC, USA, 2018.

[74] M. A. Robinson, "Using multi-item psychometric scales for research and practice in human resource management," *Human Resource Management*, vol. 57, no. 3, pp. 739–750, 2018.